Memoria PRA1 Tipología y ciclo de vida de los datos

1.Contexto

La información recolectada se trata de los datos disponibles en la página http://books.toscrape.com/ que se trata de una página con los datos de una colección de libros.

Se ha elegido esta página porque es un buen ejemplo para empezar a trabajar de forma inicial con la librería Scrapy y con el scraping de los datos en una página.

2.Título

EL titulo elegido para el dataset es book.csv al tratarse sobre datos de la colección de libros presentes en la página.

3. Descripción del dataset

El dataset extraído se contiene datos sobre una colección de los libros presentes en la página elegido para realizar el scraping.

4. Representación gráfica

Un ejemplo de la estructura de los datos recolectados sería de la siguiente forma:

Title	Price	Stock	Stars	Description	Upc	Tax
Mesaerion: The Best Science Fiction Stories 1800-1849	£37.59	In stock (19 available)	One	Andrew Barger, award-winning author and engineer, has extensively researched forgotten journals and	e30f54cea9b38190	Â£0.00

				magazines of the early 19th century……		

5. Contenido

Donde tenemos los siguientes campos de datos:

- Title: título del libro
- Price: precio de libro
- Stock: situación de stock y la cantidad
- Description: descripción sobre el libro
- Upc: el código internacional del libro
- Tax: el precio correspondiente a la tasa

6. Propietario

No ha habido aspectos éticos/legales en este caso al tratarse de una página que se usa para practicar el scraping y el manejo de la librería Scrapy.

7.Inspiración

Se ha elegido esta página para realizar el scraping por ser un ejemplo sencillo para un primer trabajo de recolección de datos y como un manejo inicial para familiarizarse con la librería Scrapy de Python-

8. Licencia

9.Código

Para la recolección de los datos se ha utilizado la librería Scrapy de Python.

La principal dificultad ha sido la forma de poder sacar las urls de las distintas páginas existentes en la barra de navegación inferior con las distintas páginas existentes.

La forma de solventarlo ha sido mediante el uso del selector xpath y mediante el plugin Xpath Helper de Chrome, esto combinado con la herramienta de inspección de desarrollador de

Chrome, permite evaluar y comprobar los distintos selectores xpath, así como sus valores de salida. Permitiendo de sacar los selectores concretos para llegar a cada elemento de la página.

10Dataset

https://zenodo.org/record/7349208#.Y31e_HbMKUk